Sentiment Analysis of Fan Discussions During the 2022 FIFA World Cup

Presentation by Kabir Chaturvedi and Monisha Patro

Introduction

The 2022 FIFA World Cup was more than just a global sports event—it was a whirlwind of emotions shared by millions of fans worldwide.

Social media platforms buzzed with excitement, disappointment, hope, and heated debates, reflecting the highs and lows of the tournament.

Given how everyone is fully equipped with internet and have the ability to post their opinions, X, Reddit etc… had many posts which had many discussions related to this event.

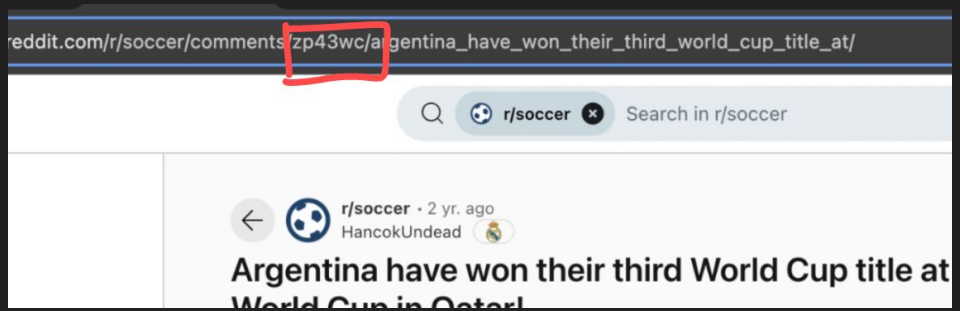
Reddit, known for its vibrant communities and unfiltered fan discussions, provided a treasure trove of raw, real-time reactions to the 2022 FIFA World Cup, making it an ideal platform for sentiment analysis.

Why did we choose this topic?

- **Sports Analytics:** Sentiment analysis in sports provides valuable insights for teams and organizations to understand fan reactions and improve engagement strategies.
- **Advertising Impact:** Brands sponsoring events like the World Cup can analyze fan sentiments to assess the effectiveness of their marketing campaigns.
- **Predictive Fan Insights:** Understanding sentiment patterns can help predict ticket sales, merchandise trends, or even viewership spikes for future events.
- **Data Exploration:** It combines sports enthusiasm with data science, showcasing the potential of natural language processing and machine learning in real-world scenarios.

Data Collection Overview

For this project, we gathered data from Reddit using its official API, focusing on fan discussions during the 2022 FIFA World Cup. Subreddits like r/soccer, r/football, and r/worldcup were selected because they hosted vibrant conversations related to the event.



Using PRAW (Python Reddit API Wrapper), we extracted post and comment data by identifying unique post IDs directly from URLs. This approach allowed us to efficiently scrape a total of 3,225 data points, ensuring our dataset captured a wide variety of perspectives.

```python
import praw
import pandas as pd
from datetime import datetime

# Initialize Reddit API client using PRAW
reddit = praw.Reddit(
    client_id='mqaw3qYOQqtDWnvUHFiUKg',
    client_secret='6bPat7wSc1DU5TN0bn4L30_tCDHIKQ',
    user_agent='test',\
    username='mozpt',
    password='Sumn@1234'
)

# Including number of comments and sort method:
submissions_info = {
    'zyucbm': {
        'start': datetime(2022, 12, 29),
        'end': datetime(2023, 4, 30),
        'category': 'post',
        'subreddit': 'worldcup',
        'comment_limit': 172,
        'sort': 'all'
    },
    'zoz9vx': {
        'start': datetime(2022, 12, 17),
        'end': datetime(2023, 4, 30),
        'category': 'post',
        'subreddit': 'worldcup',
        'comment_limit': 700,
        'sort': 'top'
    },
    '12qqii8': { # Just four months ago...
        'start': datetime(2023, 4, 17),
        'end': datetime(2024, 4, 30),
        'category': 'post',
        'subreddit': 'worldcup',
        'comment_limit': 426,
        'sort': 'all'
```

```python
    # Filter by timeframe
    for comment in selected_comments:
        comment_time = datetime.utcfromtimestamp(comment.created_utc)
        if start_date <= comment_time <= end_date:
            comment_body = comment.body
            comment_author = comment.author.name if comment.author else '[deleted]'
            comment_score = comment.score
            comment_created_utc = comment.created_utc
            row = [
                sub_id,
                sub_title,
                sub_author,
                sub_created_utc,
                sub_score,
                sub_permalink,
                comment_author,
                comment_body,
                comment_score,
                comment_created_utc
            ]
            if category == 'pre':
                pre_data.append(row)
            else:
                post_data.append(row)


# Process each submission
for sid, info in submissions_info.items():
    process_submission(sid, info)

columns = [
    'Submission_ID', 'Submission_Title', 'Submission_Author', 'Submission_Created_UTC',
    'Submission_Score', 'Submission_Permalink', 'Comment_Author', 'Comment_Body',
    'Comment_Score', 'Comment_Created_UTC'
]

pre_df = pd.DataFrame(pre_data, columns=columns)
post_df = pd.DataFrame(post_data, columns=columns)
```

Overcoming Bias in Data Collection

One major challenge in creating the dataset was avoiding bias in sentiment analysis, particularly since fan emotions post-World Cup were heavily influenced by Argentina's victory. To address this, we decided to create two separate datasets:

- **Pre-World Cup Dataset:** This dataset includes posts and comments from April 2022 to December 10, 2022, capturing the anticipation, predictions, and early discussions leading up to the event.
- **Post-World Cup Dataset:** This dataset spans from December 10, 2022, to May 2023, focusing on reactions, celebrations, and post-event reflections.

By merging these two datasets equally, we ensured a balanced representation of sentiments from both before and after the World Cup, providing a broader and less biased view of fan discussions. The final dataset of 3,225 data points reflects this thoughtful approach to data collection.

What does our dataset look like?

Dataset Columns and Their Descriptions:

1. **Submission_ID:** Unique identifier for each Reddit post, crucial for tracking and referencing posts.
2. **Submission_Title:** The title of the post, providing context and the main topic of discussion.
3. **Submission_Author:** The username of the Redditor who created the post, helpful for understanding user engagement.
4. **Submission_Created_UTC:** The timestamp (in UTC) when the post was created, useful for temporal analysis.
5. **Submission_Score:** The total number of upvotes minus downvotes on the post, indicating its popularity.

	Submission_ID	Submission_Title	Submission_Author	Submission_Created_UTC	Submission_Score
0	ttwab9	The stage is set for Qatar!	blackhole2005	1648833832	450
1	ttwab9	The stage is set for Qatar!	blackhole2005	1648833832	450
2	ttwab9	The stage is set for Qatar!	blackhole2005	1648833832	450
3	ttwab9	The stage is set for Qatar!	blackhole2005	1648833832	450
4	ttwab9	The stage is set for Qatar!	blackhole2005	1648833832	450

Submission_Permalink	Comment_Author	Comment_Body	Comment_Score	Comment_Created_UTC
https://www.reddit.com/r/worldcup/comments/ttw...	[deleted]	How is no one talking about the epic battle th...	56	1648854978
https://www.reddit.com/r/worldcup/comments/ttw...	Crabbyrob	It's still so wild to see Canada in there. Gre...	25	1648839543
https://www.reddit.com/r/worldcup/comments/ttw...	TheDickheadNextDoor	I'm reckoning Wales, Costa rica and Peru will ...	23	1648839831
https://www.reddit.com/r/worldcup/comments/ttw...	ForgingIron	It's still surreal seeing Canada here, among t...	42	1648853098
https://www.reddit.com/r/worldcup/comments/ttw...	UnpopularTruthDude	Where is italy? 🤔	21	1648841483

6. **Submission_Permalink:** A URL link to the specific Reddit post, useful for referencing back to the original discussion.
7. **Comment_Author:** The username of the Redditor who made the comment, providing insights into active contributors.
8. **Comment_Body:** The text of the comment, which is the main input for sentiment analysis.
9. **Comment_Score:** The upvote score of a comment, indicating its relevance or popularity in the discussion.
10. **Comment_Created_UTC:** The timestamp (in UTC) when the comment was created, allowing analysis of discussion timelines.

Why These Columns Were Chosen

- **Focus on Sentiment:** The combination of submission titles and comment bodies offers a robust input for sentiment analysis.
- **Popularity Indicators:** Submission and comment scores provide a measure of engagement and relevance.
- **Temporal Analysis:** Creation timestamps enable tracking of sentiment trends during key moments of the World Cup.
- **Comprehensive Context:** Including post and comment metadata ensures a complete understanding of the discussions.
- **Data Diversity:** Collecting from multiple subreddits and post types broadens the scope of analysis.

Data Cleaning Process

To ensure our sentiment analysis is accurate and meaningful, we focused on cleaning the **Comment_Body** column. Here's how we approached it:

1. **Lowercasing Text:** All comments were converted to lowercase to maintain consistency and avoid case-related mismatches.
    - Example: "THIS is Exciting!" becomes "this is exciting".
2. **Removing Unnecessary Elements:** URLs, punctuation, and special characters were removed to simplify the text and reduce noise.
    - Example: "Check this out: http://example.com!!" becomes "check this out".
3. **Trimming Extra Whitespace:** We ensured no extra spaces existed within or around the comments for uniform formatting.
4. **Removing Stopwords:** Commonly used words like "the," "is," or "and" that do not contribute to sentiment were removed using the NLTK library.
    - Example: "this is a great match" becomes "great match".

This is how the first 5 rows of our output looked like:

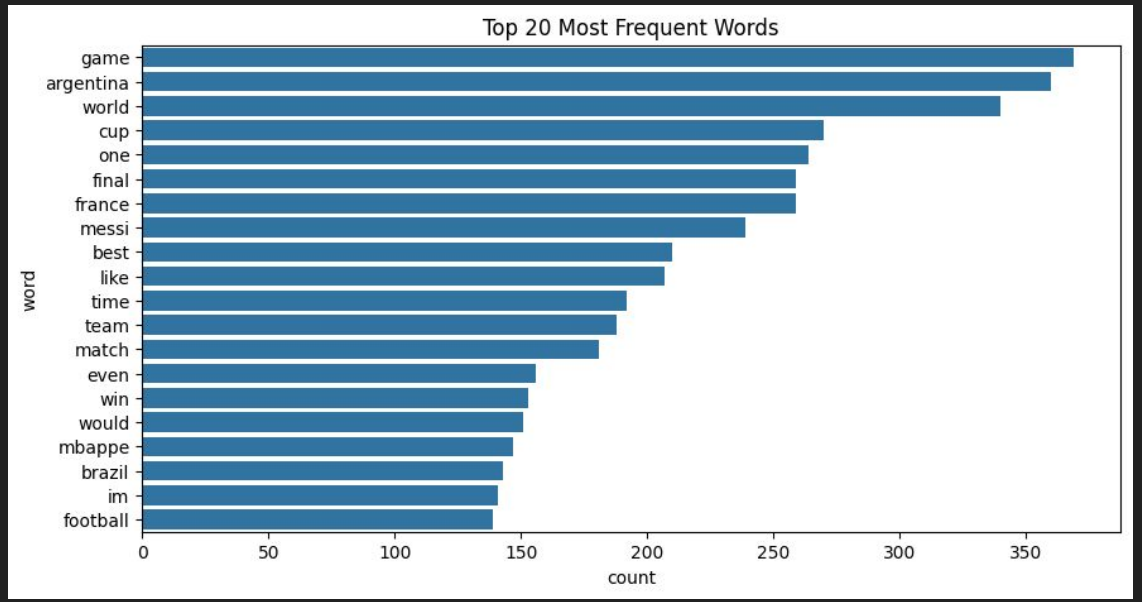
	Comment_Body	cleaned_comment
0	How is no one talking about the epic battle th...	one talking epic battle going iran vs usa
1	It's still so wild to see Canada in there. Gre...	still wild see canada great week us
2	I'm reckoning Wales, Costa rica and Peru will ...	im reckoning wales costa rica peru take last t...
3	It's still surreal seeing Canada here, among t...	still surreal seeing canada among reigning sil...
4	Where is italy? 🤔	italy

Exploratory Data Analysis (EDA) of Our Dataset

Word Cloud of Most Common Words in Comments

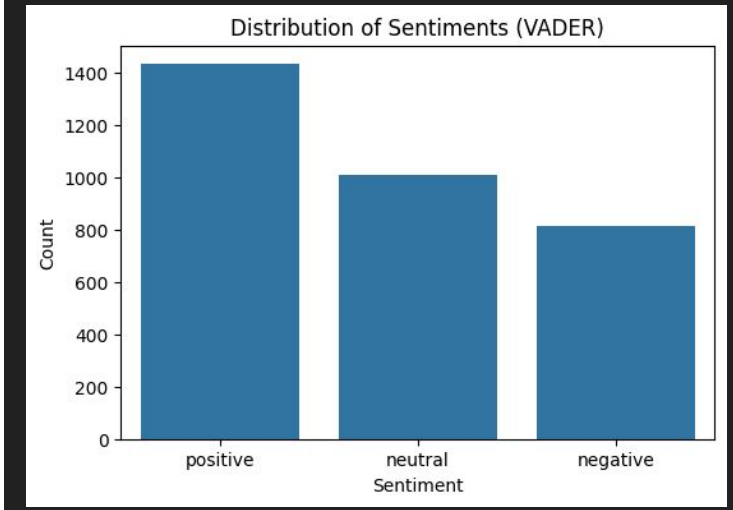
This word cloud highlights the most frequently used words in Reddit discussions about the 2022 FIFA World Cup.

Top 20 Most Frequent Words



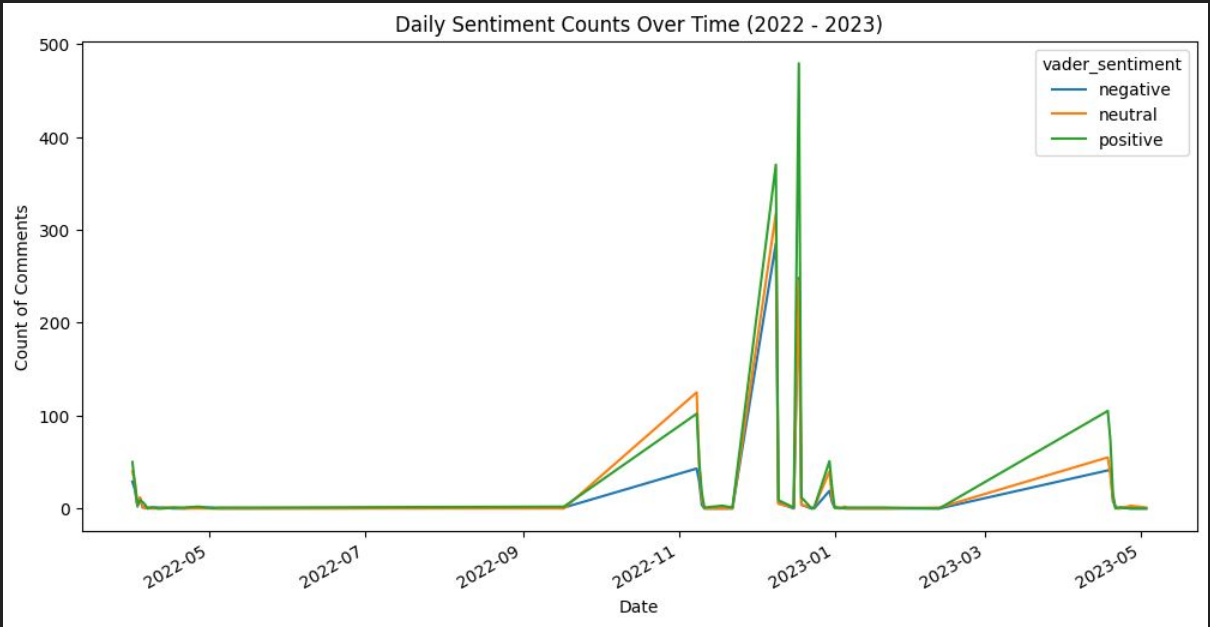
The bar chart showcases the top 20 most commonly used words in the dataset. Words like "game," "Argentina," and "world" rank highest, reaffirming their centrality in the discussions.

Distribution of Sentiments (VADER)



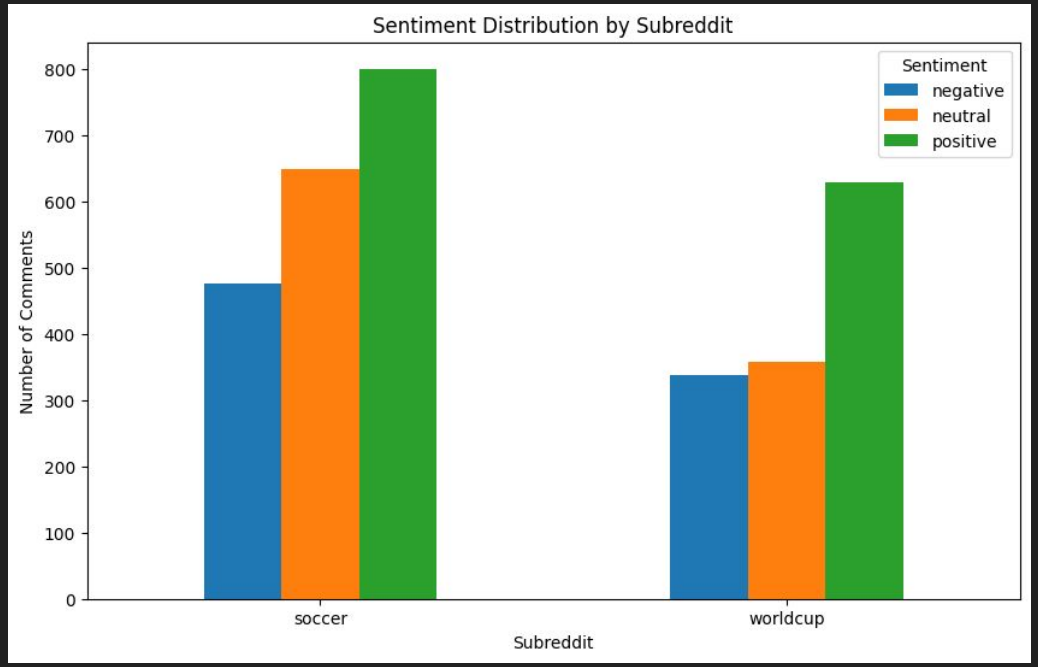
This bar chart illustrates the distribution of positive, neutral, and negative sentiments across all comments.

Daily Sentiment Counts Over Time (2022-2023)



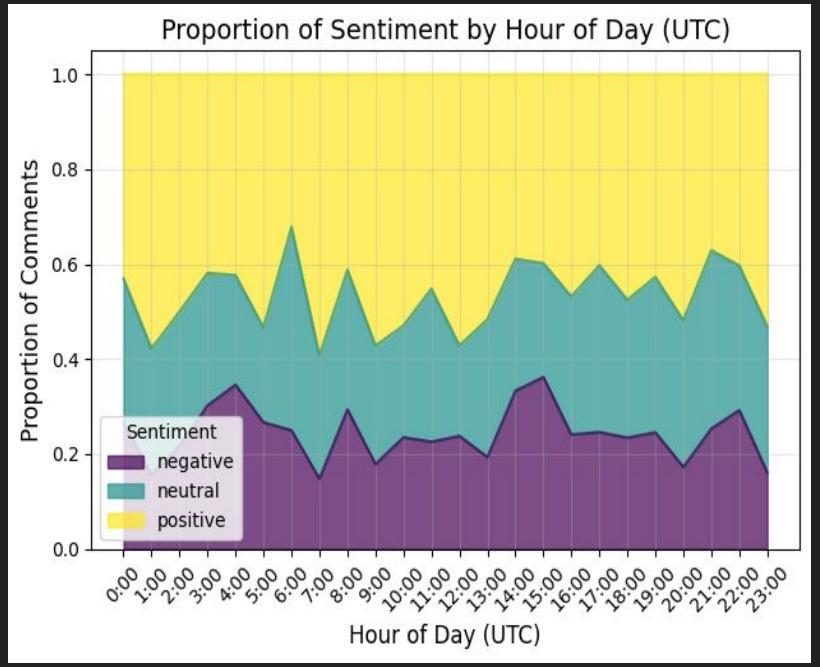
This line chart tracks sentiment trends over time, with spikes corresponding to key World Cup events, like matches or final outcomes.

Sentiment Distribution by Subreddit



This bar chart compares sentiment distribution across subreddits like r/soccer and r/worldcup.

Proportion of Sentiment by Hour of Day (UTC)



This stacked area chart shows the proportion of sentiments at different hours of the day. Positive sentiments remain consistently high throughout the day, while neutral and negative sentiments fluctuate.

Model Comparisons

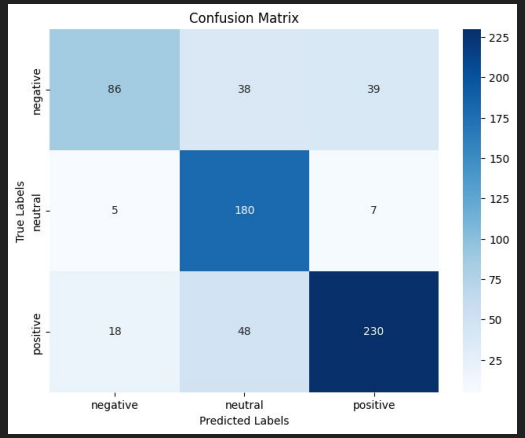
To classify fan sentiments during the World Cup, we explored multiple machine learning models and techniques.

- Used techniques like text-based feature extraction with CountVectorizer and embedding-based methods with Gensim's Word2Vec.
- Integrated advanced transformer-based models like RoBERTa for state-of-the-art performance.
- Focused on predicting sentiment labels (positive, neutral, negative) effectively.
- Analyzed and compared the strengths and limitations of these diverse approaches for a comprehensive understanding.

Comparison of Models and Findings

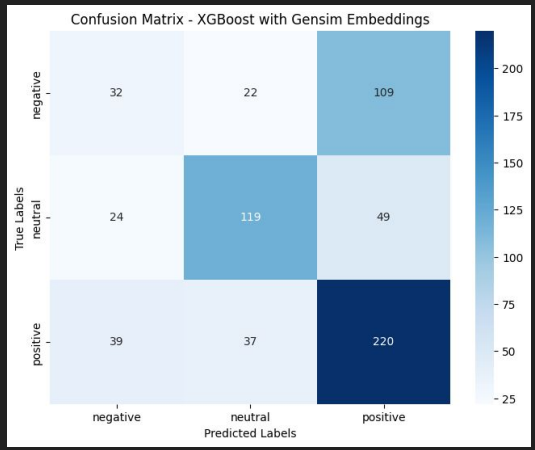
Model	Feature Extraction	Accuracy	Key Findings
XGBoost with CountVectorizer	CountVectorizer	76%	Most accurate; captured clear patterns from frequent words. Performed well for positive and neutral sentiments.
XGBoost with Word2Vec	CountVectorizer	57%	Captured semantic relationships but underperformed due to limited vocabulary and dataset size.
Random Forest with CountVectorizer	CountVectorizer	74%	Balanced approach; excelled in identifying neutral sentiments but slightly weaker for negatives.
Random Forest with Word2Vec	Word2Vec Embeddings	57%	Struggled to leverage embeddings effectively, leading to poor performance, especially for negative sentiments.

We compared models and feature extraction techniques to identify the best approach for accurately classifying fan sentiments in our dataset.



Key Takeaways

- **Best Performing Model:** XGBoost with CountVectorizer emerged as the most effective for this dataset, balancing precision and recall.
- **CountVectorizer Strength:** Simplicity and frequency-based features proved reliable for smaller datasets.
- **Challenges with Word2Vec:** Embeddings offered richer context but required more data for significant performance gains.
- **Next Steps:** Larger datasets and fine-tuned embeddings could unlock further insights, especially for nuanced sentiment analysis.



RoBERTa

After exploring traditional techniques like CountVectorizer and Word2Vec for feature extraction, and testing models like XGBoost and Random Forest, we recognized their limitations in fully understanding the nuances of fan discussions.

These methods struggled to capture deeper contextual and semantic relationships in the text. RoBERTa, a transformer-based state-of-the-art model, was chosen next for its ability to handle complex linguistic structures and deliver advanced sentiment classification.

RoBERTa because:

Twitter RoBERTa-base Sentiment Model is fine-tuned for social media text, aligning well with Reddit's informal and expressive language.

Ideal for analyzing Reddit comments with nuanced fan discussions.

Goal:
Leverage RoBERTa's pre-trained capabilities for sentiment classification.

RoBERTa Model Inference

To test RoBERTa's capabilities out of the box, we used the pre-trained Twitter RoBERTa sentiment model to classify sentiments in our dataset.

- **Steps Taken:**
  - Loaded the pre-trained RoBERTa model and tokenizer.
  - Applied the model to all cleaned comments in the dataset.
  - Mapped predictions to sentiment labels: **positive, neutral, negative.**
- **Findings:**
  - RoBERTa performed well in identifying nuanced sentiments, often aligning with human intuition.
  - For example, discussions about Argentina’s victory were correctly labeled as positive, while debates about controversial decisions were labeled as negative or neutral.

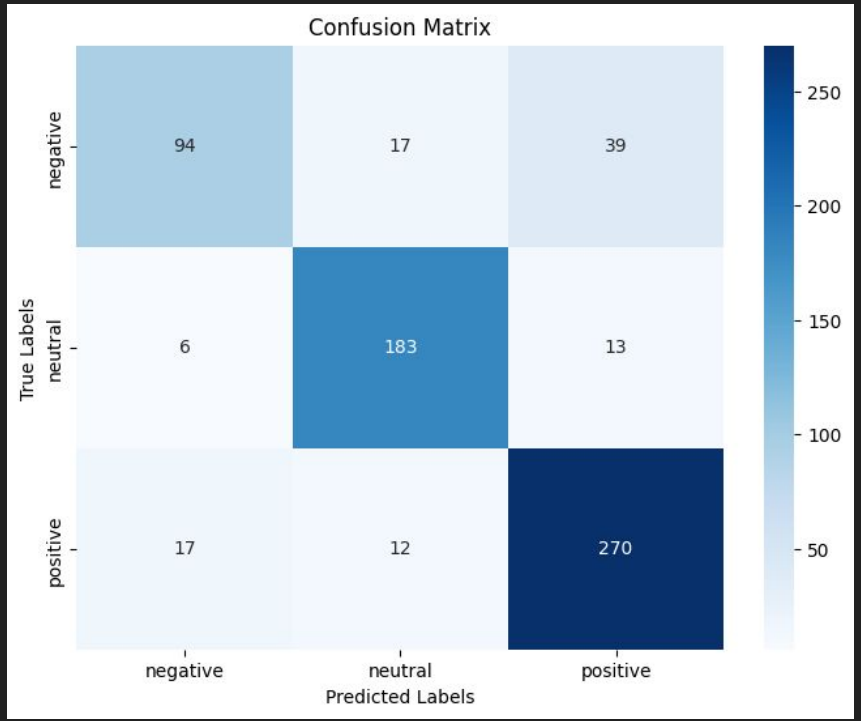
Fine-Tuning RoBERTa on Our Dataset

To further enhance performance, we fine-tuned RoBERTa on our labeled dataset using a train-test split.

- **Steps Taken:**
  - Preprocessed data by mapping labels to integers and tokenizing comments.
  - Fine-tuned RoBERTa over 2 epochs, optimizing it specifically for Reddit fan discussions.
  - Evaluated the model's performance using metrics like precision, recall, and F1-score.

Key Metrics After Fine-Tuning:

- **Accuracy:** 84%
- **Precision & Recall:** Positive and neutral sentiments were classified with high accuracy, while negative sentiments improved significantly compared to previous models.
- **Confusion Matrix:** Positive sentiments dominated correctly classified examples, followed by neutral. Some negative sentiments still overlapped with neutral or positive.



Future Steps

1. **More Data:** Fine-tune RoBERTa on a larger and more diverse dataset to further improve performance on negative sentiments.
2. **Evaluate Sports-Specific Fine-Tuned Models**: Explore the impact of using a RoBERTa model fine-tuned on sports-specific data to assess any improvements in capturing sentiments on our dataset.

References:

1. **Hugging Face Transformers**
    - *Website:* https://huggingface.co/transformers
    - Used for implementing RoBERTa models and pipelines for sentiment analysis.
2. **Cardiff NLP - Twitter Sentiment RoBERTa Model**
    - *Website:* https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
    - Pre-trained sentiment classification model used for inference and fine-tuning.
3. **Scikit-learn Documentation**
    - *Website:* https://scikit-learn.org/stable/
    - Referenced for implementing CountVectorizer, Random Forest, and XGBoost models.

4. Gensim Word2Vec

- *Website:* https://radimrehurek.com/gensim/
- Used for generating word embeddings for the dataset.

5. Kaggle - Sentiment Analysis Resources

- *Website:* https://www.kaggle.com/
- Referenced for general techniques and examples for sentiment classification.

THANK YOU